IS597MLC: Final Project Report

NetID: bhujbal3
Student Name: Neha Sunil Bhujbal

Title

Predicting Severity of Traffic Collisions in the United States

Introduction

Traffic collisions pose a significant public safety concern in the United States, leading to injuries, fatalities, and economic losses. The objective of this project is to develop a machine learning model to predict the severity of traffic collisions based on various factors such as weather conditions, road type, time of day, and others. By accurately predicting collision severity, authorities can better allocate resources, implement preventive measures, and improve emergency response. The research questions include:

1. What are the primary factors contributing to the severity of traffic collisions in the United States?
2. Can machine learning models effectively classify collision severity based on historical data?
3. How does the model's performance vary across different variants of ML algorithms?

Literature Review

Existing research has been studied in depth to understand limitations and draw inspiration to build up the proposed project architecture. Abdel-Aty and Radwan (2000) focused on modeling the occurrence and involvement of traffic accidents, aiming to identify factors contributing to accident occurrence. Utilizing statistical techniques, they analyzed accident data to identify significant variables affecting accident involvement. Their findings contribute to understanding the complex dynamics of traffic accidents and assist in developing effective preventive measures.

Quddus and Noland (2005) conducted a spatially disaggregated analysis of road casualties in England, focusing on the geographical distribution and spatial patterns of accidents. Through spatial analysis techniques, they identified hotspots and high-risk areas for road casualties, providing insights for targeted intervention strategies. Their findings contribute to understanding the spatial dynamics of road safety and inform policy-making efforts to reduce accident rates.

Wang and Zhang (2019) employed hybrid machine learning techniques to develop predictive models for traffic accident severity. By integrating multiple algorithms, including random forests and gradient boosting machines, they achieved improved accuracy in predicting accident severity. Their research highlights the effectiveness of data-driven approaches in enhancing the performance of predictive models for traffic safety analysis.

Chen and Haque (2020) conducted a comparative study to predict the severity of traffic accidents using machine learning techniques. They evaluated various algorithms, such as decision trees, support vector machines, and neural networks, to assess their effectiveness in predicting accident severity. Their research provides valuable insights into the performance of different machine learning models and their applicability in predicting traffic accident severity.

Data

A. Data Collection

The dataset used for this project is sourced from Kaggle and is titled "US Accidents". This dataset contains detailed information about traffic accidents across the United States, covering various factors such as weather conditions, road type, time of day, location coordinates, and collision severity. The dataset comprises over 3 million records with 49 features, making it suitable for training machine learning models. Each record represents a single traffic collision incident.

The dataset is provided in a CSV (Comma-Separated Values) format, consisting of multiple columns representing different attributes of the accidents. The primary target class for this project is the "Severity" column, which indicates the severity level of each collision.

To ensure that the dataset meets the requirement of having at least 30,000 instances, a subset containing 50,000 records from the original dataset has been used for the project. The subset was randomly sampled from the original dataset to maintain diversity and representativeness.

Dataset link: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

Attribute	**Description**
ID	A unique identifier for each accident
Source	Source of the accident report
TMC	Traffic Message Channel code, which provides more detailed accident description
Severity	Severity of the accident, ranging from 1 to 4 (1 being the least severe and 4 being the most severe)
Start_Time	Start time of the accident
End_Time	End time of the accident
Start_Lat	Latitude coordinate of the accident start location
Start_Lng	Longitude coordinate of the accident start location
End_Lat	Latitude coordinate of the accident end location
End_Lng	Longitude coordinate of the accident end location
Distance(mi)	Distance of the accident from the start location
Description	Description of the accident
Number	Street number where the accident occurred
Street	Street where the accident occurred
Side	Side of the street where the accident occurred (left or right)
City	City where the accident occurred
County	County where the accident occurred
State	State where the accident occurred
Zipcode	Zipcode where the accident occurred
Country	Country where the accident occurred
Timezone	Timezone of the accident location
Airport_Code	Airport code near the accident location
Weather_Timestamp	Timestamp of the weather report
Temperature(F)	Temperature in Fahrenheit at the accident location
Wind_Chill(F)	Wind chill temperature in Fahrenheit
Humidity(%)	Humidity percentage at the accident location
Pressure(in)	Atmospheric pressure in inches of mercury at the accident location
Visibility(mi)	Visibility in miles at the accident location

Wind_Direction	Wind direction at the accident location
Wind_Speed(mph)	Wind speed in miles per hour at the accident location
Precipitation(in)	Precipitation amount in inches at the accident location
Weather_Condition	Weather condition at the accident location
Amenity	Indicates whether there is an amenity near the accident location (e.g., restroom, parking)
Bump	Indicates whether there is a speed bump near the accident location
Crossing	Indicates whether there is a crossing near the accident location
Give_Way	Indicates whether there is a give way near the accident location
Junction	Indicates whether there is a junction near the accident location
No_Exit	Indicates whether there is a no exit near the accident location
Railway	Indicates whether there is a railway near the accident location
Roundabout	Indicates whether there is a roundabout near the accident location
Station	Indicates whether there is a station near the accident location
Stop	Indicates whether there is a stop near the accident location
Traffic_Calming	Indicates whether there is a traffic calming device near the accident location
Traffic_Signal	Indicates whether there is a traffic signal near the accident location
Turning_Loop	Indicates whether there is a turning loop near the accident location
Sunrise_Sunset	Indicates whether the accident occurred during sunrise or sunset
Civil_Twilight	Indicates whether the accident occurred during civil twilight
Nautical_Twilight	Indicates whether the accident occurred during nautical twilight
Astronomical_Twilight	Indicates whether the accident occurred during astronomical twilight

Image of Dataset:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	ID	Source	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Description	Street	City	County	State	Zipcode	Country	Timezone	Airport_Code	Weather_Tir	Temperature
2	A-2047758	Source2	2	##########	##########	30.641211	-91.153481			0	Accident on I	Highway 19	Zachary	East Baton R	LA	70791-4610	US	US/Central	KBTR	##########	77
3	A-4694324	Source1	2	37:14.0	56:53.0	38.990562	-77.39907	38.990037	-77.398282	0.056	Incident on F	Forest Ridge	Sterling	Loudoun	VA	20164-2813	US	US/Eastern	KIAD	##########	45
4	A-5006183	Source1	2	13:00.0	22:45.0	34.6611893	-120.49282	34.6611893	-120.49244	0.022	Accident on \	Floradale Av	Lompoc	Santa Barbai	CA	93436	US	US/Pacific	KLPC	##########	68
5	A-4237356	Source1	2	##########	##########	43.680592	-92.993317	43.680574	-92.972223	1.054	Incident on I-	14th St NW	Austin	Mower	MN	55912	US	US/Central	KAUM	##########	27
6	A-6690583	Source1	2	##########	##########	35.395484	-118.98518	35.395476	-118.986	0.046	RP ADV THEY	River Blvd	Bakersfield	Kern	CA	93305-2649	US	US/Pacific	KBFL	##########	42
7	A-1101469	Source2	2	##########	##########	42.532082	-70.944267			0	Accident on I	Lowell St	Peabody	Essex	MA	01960-4275	US	US/Eastern	KBVY	##########	42
8	A-7222249	Source1	2	##########	##########	42.42128	-123.11945	42.42128	-123.11945	0	At OR-99/Ex	I-5 N	Gold Hill	Jackson	OR	97525	US	US/Pacific	KMFR	##########	35
9	A-6198239	Source1	2	48:00.0	09:09.0	30.19101	-85.682508	30.190329	-85.68253	0.047	Incident on C	Claremont D	Panama City	Bay	FL	32405-3534	US	US/Central	KPAM	##########	90
10	A-4222549	Source1	2	##########	##########	32.868947	-96.804018	32.8695	-96.804014	0.038	Incident on P	Preston Rd	Dallas	Dallas	TX	75225	US	US/Central	KDAL	##########	91
11	A-5924038	Source1	2	##########	##########	39.71721768	-86.124691	39.73347768	-86.137021	1.301	Incident on I-	I-65	Indianapolis	Marion	IN	46237	US	US/Eastern	KIND	##########	63
12	A-925338	Source2	2	##########	##########	39.93346	-86.157433			2.480000019	Exit ramp frc	N Meridian S	Indianapolis	Hamilton	IN	46290	US	US/Eastern	KTYQ	##########	70
13	A-4908440	Source1	2	##########	##########	47.25825905	-115.05292	47.28336905	-115.07781	2.091	Travelers car	I-90 W	Saint Regis	Mineral	MT	59866	US	US/Mountai	K3TH	##########	13
14	A-1388988	Source2	2	##########	##########	34.72015	-86.616592			0	Lane blocked	Governors Di	Huntsville	Madison	AL	35805-3542	US	US/Central	KHUA	##########	85
15	A-4535214	Source1	2	##########	##########	32.771645	-117.16141	32.730856	-117.15468	2.845	Slow traffic (	Friars Rd	San Diego	San Diego	CA	92108	US	US/Pacific	KMYF	##########	63
16	A-2127689	Source2	2	##########	##########	33.436073	-111.92616			0	Right hand s	N Scottsdale	Tempe	Maricopa	AZ	85281	US	US/Mountai	KPHX	##########	64
17	A-6609749	Source1	2	##########	##########	25.89866	-80.382801	25.895972	-80.379142	0.294	Stationary tr	W Okeechol	Miami	Miami-Dade	FL	33178	US	US/Eastern	KMIA	##########	83
18	A-6214306	Source1	2	##########	##########	38.132332	-77.511383	38.12196	-77.51632	0.765	Slow traffic (	I-95 S	Woodford	Spotsylvania	VA	22580	US	US/Eastern	KEZF	##########	84
19	A-2881976	Source2	2	##########	##########	29.75239	-95.364708			0	Accident on (	Caroline St	Houston	Harris	TX	77002-6904	US	US/Central	KMCJ	##########	64.4
20	A-2635201	Source2	2	##########	##########	41.926895	-73.912605			0	Right hand s	Mill St	Rhinebeck	Dutchess	NY	12572-1427	US	US/Eastern	KPOU	##########	84
21	A-5659848	Source1	2	##########	##########	25.794969	-80.258877	25.794973	-80.25903	0.01	Incident on N	NW 21st St	Miami	Miami-Dade	FL	33142-6704	US	US/Eastern	KMIA	##########	70

	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT
mi)	Wind_Direct	Wind_Speed	Precipitation	Weather_Co	Amenity	Bump	Crossing	Give_Way	Junction	No_Exit	Railway	Roundabout	Station	Stop	Traffic_Calm	Traffic_Signa	Turning_Loo	Sunrise_Sun	Civil_Twiligh	Nautical_Tw	Astronomica
10	NW	5	0	Fair	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	Day	Day	Day	Day
10	W	5	0	Fair	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Night	Night	Night	Night
10	W	13	0	Fair	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	Day	Day	Day	Day
10	ENE	15	0	Wintry Mix	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Day	Day	Day	Day
10	CALM	0	0	Fair	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Night	Night	Night	Night
10	W	13	0	Fair	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	Day	Day	Day	Day
10	CALM	0	0	Light Rain	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Day	Day	Day	Day
10	SW	12	0	Fair	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Day	Day	Day	Day
10	VAR	7	0	Fair	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	Day	Day	Day	Day
10	SW	10	0	Cloudy	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Night	Day	Day	Day
10	S	3	0	Cloudy	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Day	Day	Day	Day
10	VAR	3	0	Cloudy	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Night	Night	Night	Night
10	S	7	0	Fair	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	Day	Day	Day	Day
10	NW	14	0	Fair	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Day	Day	Day	Day
10	E	7	0	Fair	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	Day	Day	Day	Day
10	WSW	14	0	Mostly Cloud	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Day	Day	Day	Day
10	NNE	3	0	Partly Cloudy	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Day	Day	Day	Day
9	Variable	4.6		Clear	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	Day	Day	Day	Day
10	West	5.8		Scattered Clc	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE	Day	Day	Day	Day
10	N	3	0	Mostly Cloud	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	Night	Night	Night	Night
10	NNE	6.9		Clear	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Day	Day	Day	Day

B. Data Pre-processing

The dataset underwent several pre-processing steps to ensure its quality and suitability for analysis. These steps included:

1. **Handling of Missing Values**: Due to the extensive dataset, features containing null values exceeding 40% were eliminated to enhance prediction accuracy.
2. **Duplicate Removal**: Checked for duplicates in the dataset and dropped any duplicate rows using the **drop_duplicates()** function.
3. **Categorical Variable Conversion**: Categorical columns were converted into numerical format using one-hot encoding or label encoding techniques, depending on the nature of the variables and the requirements of the machine learning algorithms.
4. **Train-Test Split**: The dataset was split into training and testing sets with an 80:20 ratio to facilitate model training and evaluation.
5. **Dimensionality Reduction**: Given the large number of features, Principal Component Analysis (PCA) was employed for dimensionality reduction. PCA helped select the optimal features that explained the most variance in the data, thereby improving computational efficiency and reducing overfitting risks.

Overall, these pre-processing steps were crucial for ensuring the quality, integrity, and efficiency of the dataset for subsequent analysis and model development. The combination of pandas, NumPy, scikit-learn, and other libraries facilitated effective handling of duplicates, missing values, categorical variables, and dimensionality reduction tasks.

Methodology

The objective of this study has been to develop a predictive model capable of accurately classifying the severity of traffic accidents based on diverse attributes present in the dataset. To achieve this goal, a systematic approach has been undertaken, encompassing several essential steps.

Firstly, exploratory data analysis (EDA) has been conducted to delve into the distribution and relationships among different features. Utilizing visualization techniques such as histograms, scatter plots, and correlation matrices, patterns and correlations within the dataset have been identified. Additionally, descriptive statistics have been employed to succinctly summarize the key characteristics of the dataset, laying a robust foundation for subsequent analyses.

Next, a comprehensive array of machine learning algorithms has been explored to train predictive models. These algorithms, including logistic regression, decision trees, random forests, support vector machines, and gradient boosting, have been meticulously implemented utilizing libraries such as scikit-learn in Python. Each algorithm has been rigorously evaluated for its effectiveness in accurately predicting traffic accident severity, considering factors such as model performance, computational efficiency, and interpretability.

Results

The results of Exploratory Data Analysis reveal several factors that are responsible for the severity of accidents. These include 'State', 'Visibility(mi)', 'Wind_Speed(mph)', 'Precipitation(in)', 'Amenity', 'Bump', 'Crossing', 'Give_Way', 'Junction', 'No_Exit', and 'Sunrise_Sunset'.

Next, the effectiveness of machine learning models in classifying collision severity based on historical data has been explored through the evaluation of four different algorithms: Logistic Regression, Support Vector Machine (SVM), Random Forests, and Adaboost. The results of the classification performance metrics are summarized as follows:

Model	Accuracy	Precision	Recall	F-1 Score
Logistic Regression	0.57	0.54	0.565	0.542
Support Vector Machine	0.61	0.6	0.605	0.572
Random Forests	0.92	0.922	0.92	0.91
Adaboost	0.54	0.51	0.537	0.495

The results demonstrate notable variations in performance across the different models.

1. **Logistic Regression and Adaboost**: These models exhibit lower performance compared to others, with accuracy scores of 0.57 and 0.54, respectively. While Logistic Regression and Adaboost provide reasonable precision and recall scores, their overall classification accuracy and F-1 scores are comparatively lower. This suggests that they may struggle with effectively capturing the complexity of the data and distinguishing between different severity levels.

2. **Support Vector Machine (SVM)**: SVM demonstrates slightly improved performance with an accuracy score of 0.61. While SVM achieves higher precision and recall scores compared to Logistic Regression and Adaboost, its overall performance still falls short when compared to Random Forests

3. **Random Forests**: Random Forests outperform the other models significantly, achieving an accuracy score of 0.92. This high accuracy is accompanied by impressive precision, recall, and F-1 scores, indicating robust classification performance. Random Forests leverage ensemble learning to combine multiple decision trees, allowing for a more sophisticated and accurate classification of collision severity based on historical data.

Discussion & Conclusion

The observed variations in model performance can be attributed to several factors. Random Forests excel in handling high-dimensional data and capturing complex relationships between features, making them well-suited for classification tasks involving historical data with multiple attributes. In contrast, Logistic Regression, Adaboost, and SVM may struggle with non-linear relationships and feature interactions present in the data, leading to inferior performance.

Additionally, the imbalance in the distribution of severity levels within the dataset may impact model performance. Random Forests, with their inherent ability to handle class imbalances, are less affected by this issue compared to other models.

In conclusion, the results indicate that Random Forests are highly effective in classifying collision severity based on historical data, outperforming Logistic Regression, Support Vector Machine, and Adaboost. However, further exploration and refinement of model parameters, feature engineering techniques, and data preprocessing methods may offer opportunities for enhancing the performance of all models in future studies.

GitHub Repo

https://github.com/nehabhujbalillini/IS597_MLC_FinalProject

For simplicity, the GitHub repo contains only one Jupyter notebook, which needs to be run in order to reproduce the results. The "Data" folder needs to be placed in the same folder as the Jupyter Notebook.

References

- Chen, H., & Haque, M. M. (2020). Predicting the Severity of Traffic Accidents Using Machine Learning Techniques: A Comparative Study. *Transportation Research Part C: Emerging Technologies*, 91, 77-93.
- Abdel-Aty, M., & Radwan, A. E. (2000). Modeling Traffic Accident Occurrence and Involvement. *Accident Analysis & Prevention*, 32(5), 633-642.
- Wang, Y., & Zhang, X. (2019). Data-Driven Predictive Modeling of Traffic Accident Severity Using Hybrid Machine Learning Techniques. *Transportation Research Part C: Emerging Technologies*, 99, 212-226.
- Quddus, M. A., & Noland, R. B. (2005). A Spatially disaggregate analysis of road casualties in England. *Accident Analysis & Prevention*, 37(1), 73-81.
- Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A Countrywide Traffic Accident Dataset.
- Moosavi, S., Samavatian, M. H., Parthasarathy, S., Teodorescu, R., & Ramnath, R. (2019). Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights. *In Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*.